CoRT-Predictor: Chain of Risk Thought Autoregressive Trajectory Predictor for Autonomous Driving

Supplementary Material

1. Implementation Details of Preprocessing

Due to space constraints, detailed discussions of the Ego Data, Tokenizer, and Embedder designs are skipped in the main text, as these designs have been well established and commonly adopted in previous Transformer studies. However, since these designs largely ensure the performance and robustness of the CoRT-Predictor, we provide supplementary details regarding their configurations herein.

1.1. Ego Data

Our model is configured to use 8 frames of historical data as the condition input. The ego data consist of the vehicle's past 8 frames of $[x, y, z, L, W, H, \text{yaw}]$. Upon input to the model, the relative displacement $[\Delta x, \Delta y]$ between consecutive frames is computed based on the $[x, y]$ coordinates of each frame, which is then further represented in polar coordinates $[\theta, r]$. Subsequently, the instantaneous velocity v of the ego vehicle at each frame is computed from the positional data of the preceding and succeeding frames. The resulting ego vehicle data for that frame is then represented as $[\theta, r, v]$. Simultaneously, the model's predicted trajectory is represented in the form of relative displacement polar coordinates $[\theta, r, v]$ with respect to the previous frame. After loss computation, these predictions are converted back to Cartesian coordinates $[x, y]$ based on the position of the last historical frame. Finally, the autoregressively predicted future trajectory comprising 32 frames is collectively processed by the trajectory smoother for refinement and smoothing.

2. Other participants data

To autoregressively predict the ego vehicle's future multi-frame trajectory, it is necessary to incorporate the future positions of other participants and compute the corresponding Polar Risk Spectrum for vehicles at future time steps. To reduce model complexity, we pretrain a Transformer composed solely of temporal causal autoregressive components to infer the short-term displacement changes of other participants. When predicting the future positions of other participants, To reduce model complexity, we pretrain a network composed solely of temporal causal autoregressive components to infer the short-term displacement changes of other participants. During prediction, the model input continues to consist of the processed polar coordinate representations of these participants. To reduce model complexity, we pretrain a Transformer composed solely of temporal causal autoregressive components to infer the short-term displacement changes of other participants. During prediction, the model input continues to consist of the processed polar coordinate representations of these participants. After obtaining a vehicle's displacement in polar coordinates, we convert it back to Cartesian coordinates $[x, y]$ for the purpose of computing the Polar Risk Spectrum for future frames.

2.1. Discretization Tokenizer

Each data element within the processed other vehicle, lane Polar Risk Spectrum, and ego data undergoes normalization. This normalization is performed using Min-Max scaling, also referred to as linear scaling: $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$

Subsequently, the normalized value x_{norm}, constrained within the range $[0, 1]$, is assigned to a discrete interval. This discretization divides the interval $[0, 1]$ into N segments, and x_{norm} is allocated to the i-th segment such that $x_{\text{norm}} \approx \frac{i}{N}$, where $i, N \in Z$. In our CoRT-Predictor, N is set to 512. Following discretization, the model only needs to learn a finite set of token embeddings, making the discrete tokens more compatible with the Transformer's sequential modeling mechanism. Moreover, minor fluctuations in continuous values are mapped to the same discrete interval, preserving the token representation and enhancing the model's robustness to noise.

2.2. Fourier Embedder

We have normalized the data to the range $[0, 1]$ using a Discretization Tokenizer, such that $x_{\text{norm}} \approx \frac{i}{N}$. The normalized data is then decomposed into N_F Fourier series components and passed through a dedicated MLP to obtain the corresponding embedding Emb. By reprojecting through Fourier embedding into a continuous space, the model can recover smooth geometric structures within its internal representations, facilitating the processing of continuously varying trajectories and risk information. This approach combines the advantages of classification-based supervision (for stable learning) and continuous-space reasoning (for fine-grained modeling), where discretization serves as the "encoding" step and Fourier embedding functions as the "decoding and enhancement" process.

3. Additional Experimental Results

To further validate the generalization ability of our method on large-scale public datasets, we provide additional experimental results on the Waymo Open Motion Dataset

Table 1. Additional experimental results on the WOMD and ablation studies results.

Leaderboard, test split Method	RMM ↑	Kinematic metrics↑	Interactive metrics↑	Map-based metrics↑	minADE↓	model params↓
GUMP	0.743	0.478	0.789	0.836	1.604	523M
SMART	0.751	0.445	0.805	0.857	1.545	8M
GameFormer	0.752	0.436	0.804	0.864	1.375	15M
KiGRAS	0.760	0.469	0.806	0.866	1.438	1M
CarPlanner	0.764	0.465	0.811	0.869	1.419	13M
Ablation A	0.755	0.468	0.771	0.823	**1.356**	11M
Ablation B	0.683	0.415	0.726	0.779	1.520	5M
Ablation C	0.755	0.446	0.807	0.866	1.437	5M
Ours	**0.769**	**0.483**	**0.815**	**0.873**	1.359	5M

(WOMD). Compared to RFSD, which focuses on emergency scenarios, WOMD primarily consists of normal driving conditions with diverse multi-agent interactions. This complementary evaluation helps assess whether the proposed method generalizes beyond high-risk scenarios.

3.1. Public Benchmark Evaluation on WOMD

We conduct experiments on WOMD following the standard open-loop behavior cloning (BC) protocol, where all compared methods are trained for 32 epochs under the same training and evaluation settings. Results are reported on the official test split leaderboard, as shown in Table 1.

As observed, our method achieves the best performance across all evaluated metrics, demonstrating its effectiveness on large-scale real-world datasets. Notably, despite its lightweight architecture, our model outperforms stronger baselines with significantly larger parameter counts, indicating its efficiency in modeling trajectory prediction.

These results suggest that the proposed Polar Risk Spectrum (PRS) representation does not lead to underfitting in normal driving scenarios. Instead, by suppressing irrelevant agents that do not affect the ego vehicle's motion, the model reduces decision noise and produces more stable and accurate predictions.

3.2. Ablation Study on PRS Design

We further conduct ablation studies on WOMD to analyze the effectiveness of the proposed PRS, as summarized in Table 1.

First, a direct comparison is provided in Table 1 (Ablation A), where the PRS is replaced by a standard learned feature representation. As shown, the learned-feature baseline does not outperform our PRS-based model. In addition, the learned-feature design introduces multiple self-attention and cross-attention modules to encode surrounding context, resulting in a substantially larger parameter count. Based on our empirical observations, such learned representations tend to attend to agents that are irrelevant to the ego vehicle, thereby introducing decision noise. In contrast, the PRS explicitly suppresses the influence of irrelevant agents, leading to more stable and effective planning performance.

Second, we investigate the impact of PRS resolution by varying the number of PRS sectors. Ablation B adopts a low-resolution PRS with 4 sectors, while Ablation C uses an 12-sector configuration. While a larger number of sectors can be advantageous in rare cases such as crowded scenes or complex intersections, we observe that the 8-sector PRS (Ours) achieves the best overall generalization. We attribute this to the limited and non-uniform coverage of rare scenarios in the training data: a moderate sector granularity allows the model to more reliably learn the semantics of each sector and results in more robust planning.

These results demonstrate that the proposed PRS design provides an effective balance between representation capacity and generalization ability.

4. Visual Ablation Study

4.1. Intension Token (IT)

As shown in Figure 1, enabling the Intention Token allows the predicted ego vehicle trajectory to more fully complete a maneuver, rather than prematurely aborting it immediately after avoiding collision risks.

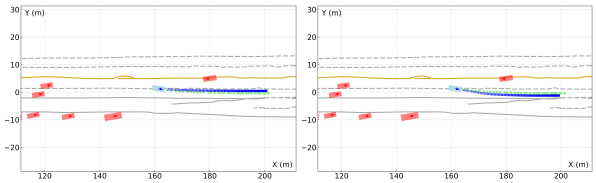


Figure 1. Visual results: Left: without IT; Right: with IT.

4.2. Trajectory Smoother (TS)

As shown in Figure 2, the Trajectory Smoother effectively smooths the ego vehicle's trajectory, reducing oscillations. This enhancement brings the predicted trajectory closer to the smooth expert trajectory.

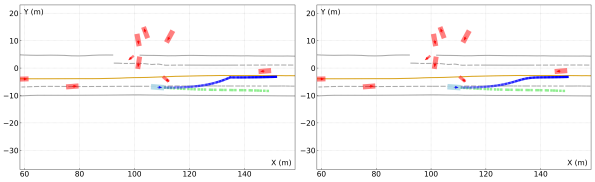
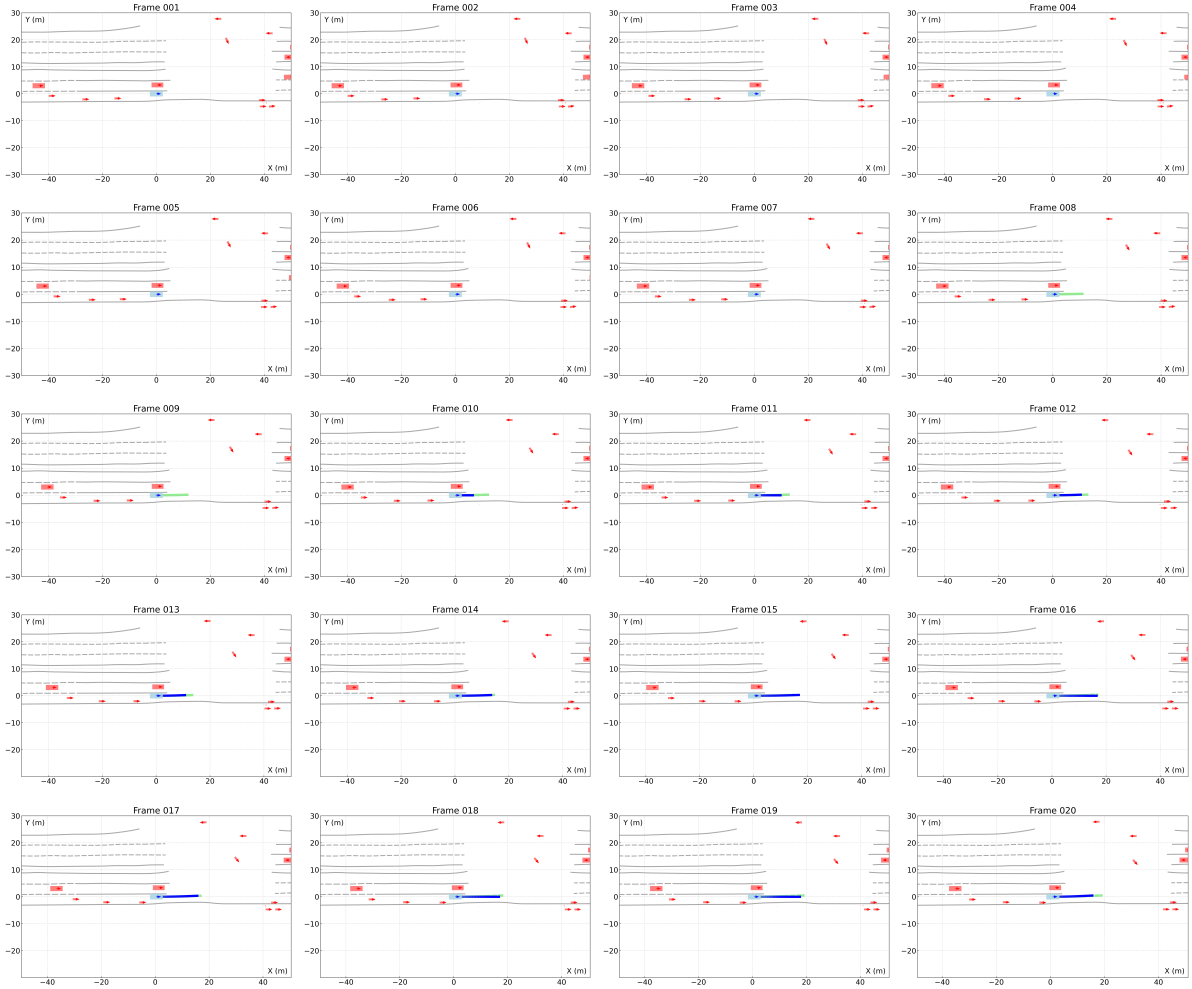
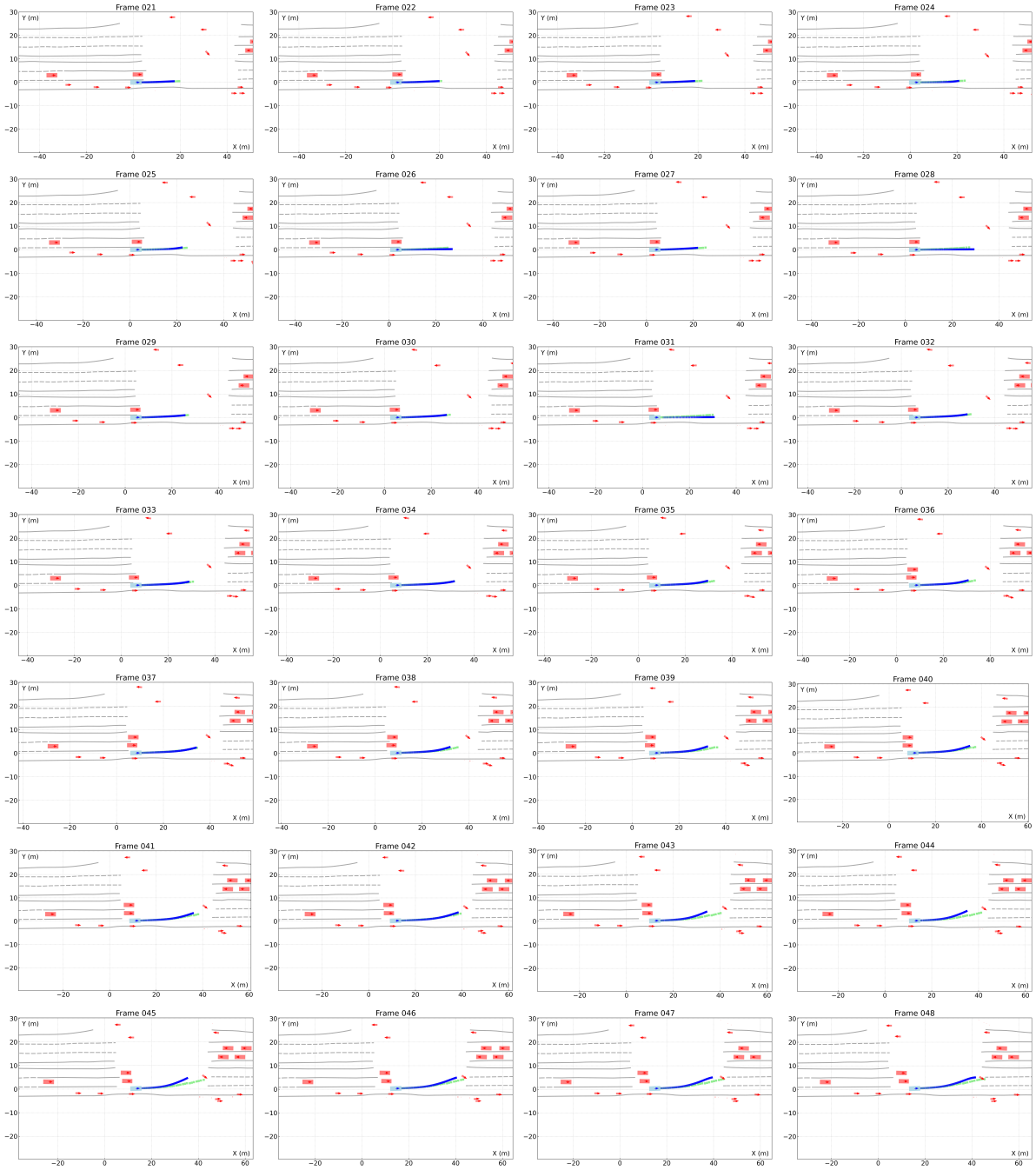


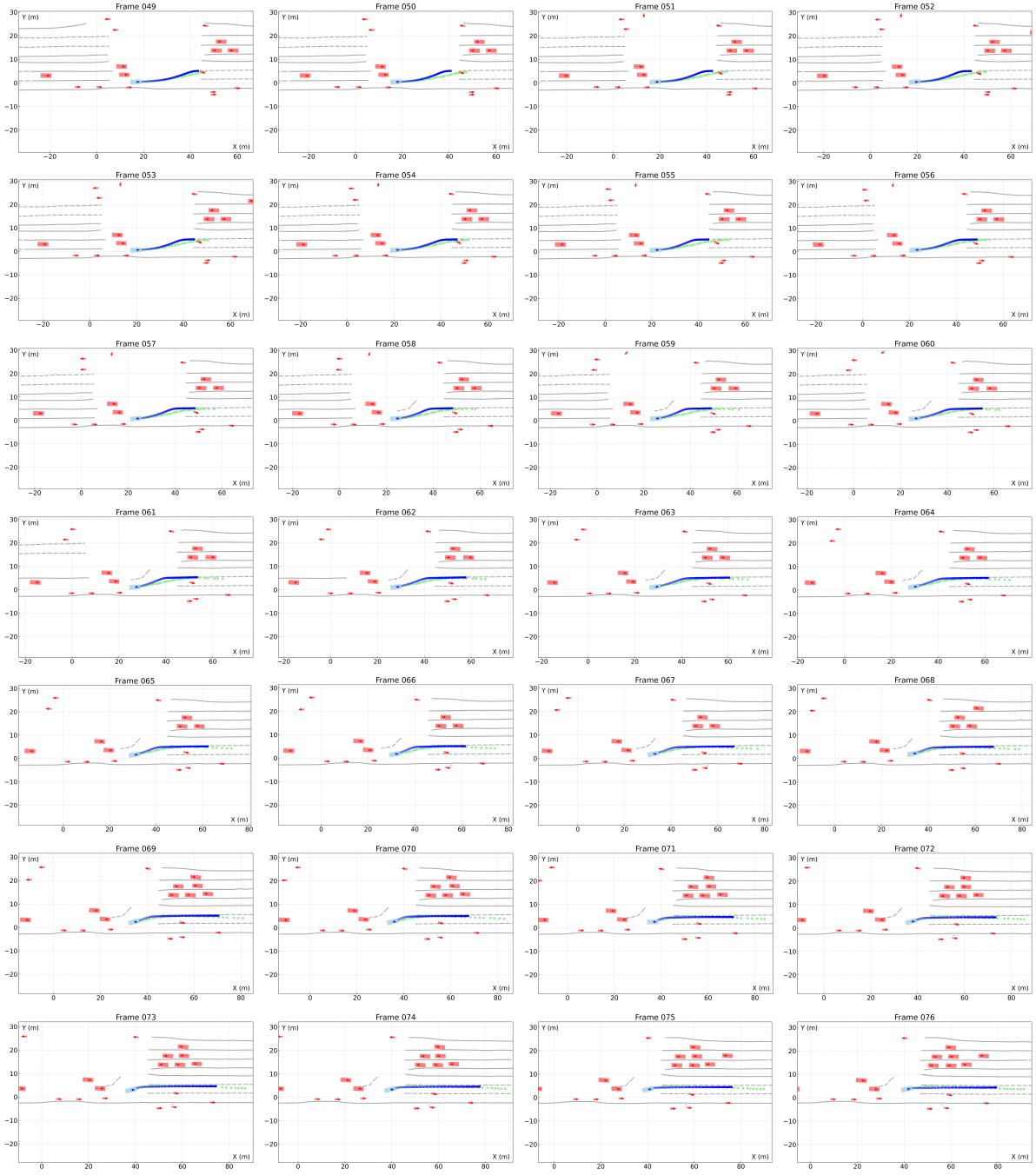
Figure 2. Visual results: Left: without TS; Right: with TS.

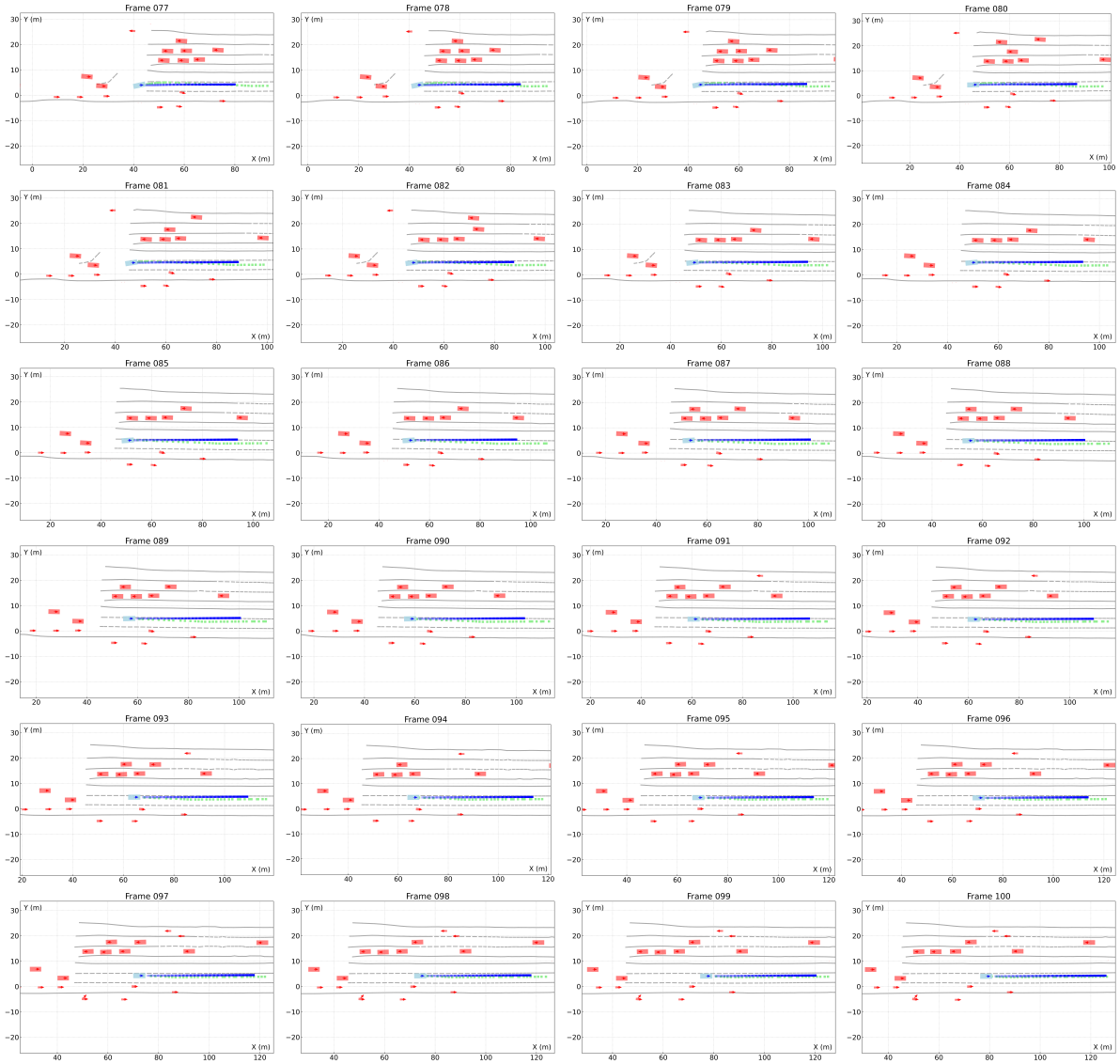
5. Frame-by-Frame Visualizations

Below, we present frame-by-frame visualizations of the inference process for a complex case. This case includes scenarios of vehicle start and acceleration (frames 1–10), intersection crossing (frames 20–80), encountering a bicycle followed by active front-wheel steering avoidance (frames 20–80), and straight driving (frames 80–100). In this scenario, our CoRT-Predictor demonstrates excellent performance and robustness when confronting multiple complex risks, including those posed by non-motorized vehicles. In the figure, the blue markers represent the ego vehicle, while the red markers denote other traffic participants. The blue lines indicate the predicted trajectory of the ego vehicle, and the green points correspond to the ground-truth future trajectory.


Frame 001
Frame 002
Frame 003
Frame 004
Frame 005
Frame 006
Frame 007
Frame 008
Frame 009
Frame 010
Frame 011
Frame 012
Frame 013
Frame 014
Frame 015
Frame 016
Frame 017
Frame 018
Frame 019
Frame 020
Y (m)
X (m)


Frame 021
Frame 022
Frame 023
Frame 024
Frame 025
Frame 026
Frame 027
Frame 028
Frame 029
Frame 030
Frame 031
Frame 032
Frame 033
Frame 034
Frame 035
Frame 036
Frame 037
Frame 038
Frame 039
Frame 040
Frame 041
Frame 042
Frame 043
Frame 044
Frame 045
Frame 046
Frame 047
Frame 048
Y (m)
X (m)


Frame 049
Frame 050
Frame 051
Frame 052
Frame 053
Frame 054
Frame 055
Frame 056
Frame 057
Frame 058
Frame 059
Frame 060
Frame 061
Frame 062
Frame 063
Frame 064
Frame 065
Frame 066
Frame 067
Frame 068
Frame 069
Frame 070
Frame 071
Frame 072
Frame 073
Frame 074
Frame 075
Frame 076
Y (m)
X (m)


Frame 077
Frame 078
Frame 079
Frame 080
Frame 081
Frame 082
Frame 083
Frame 084
Frame 085
Frame 086
Frame 087
Frame 088
Frame 089
Frame 090
Frame 091
Frame 092
Frame 093
Frame 094
Frame 095
Frame 096
Frame 097
Frame 098
Frame 099
Frame 100
Y (m)
X (m)